Supplementary Materials

From Co-expression to Co-regulation: how many microarray experiments do we need?

Ka Yee Yeung*, Mario Medvedovic[‡], Roger E. Bumgarner*

*Dept of Microbiology, Box 358070, University of Washington, Seattle, WA 98195
[‡]Center for Genome Information, Department of Environmental Helath, University of Cincinnati Medical Center, Cincinnati, OH 45267

Table of Contents

A. Effect of clustering algorithms

MCLUST vs. IMM

Our previous work (Yeung et al. 2003; Medvedovic et al. 2004) showed that IMM produced higher quality clusters relative to other algorithms (as measured by the accuracy of assigning objects to their true underlying groups on synthetic datasets with repeated measurements). To our surprise, IMM produces clustering results with relatively fewer co-regulated genes than the equal-volume spherical model from MCLUST. The IMM results shown in Figures 2 to 6 in the main manuscript are generated using the *unequal*-volume spherical model, while the MCLUST results are generated using the *equal*-volume spherical model. Since the equal-volume spherical model in MCLUST produces significantly higher z-scores than the unequal-volume spherical model in MCLUST (see Figure A.1.b from the Supplementary Materials), we extended IMM to include the equal-volume spherical model so as to investigate the reasons that lead to the relatively low proportion of co-regulated genes. Our results showed that the equal-volume spherical model of IMM does *not* significantly increase the level of co-regulation to that from the corresponding model in MCLUST (see Figure A.1.c in Supplementary Materials). Furthermore, we observed that the EM step in MCLUST does not significantly improve the z-scores after the hierarchical initialization step (Figures A.1.b and E.1.a from the Supplementary Materials). Since IMM is initialized in a randomized manner, we hypothesized that the hierarchical initialization step in MCLUST is crucial in its relatively high performance. Both the expectation-maximization (EM) algorithm employed by MCLUST and the Gibbs sampler employed by IMM are likely to run into convergence problems with high-dimensional data due to potentially large numbers of local maxima in the case of MCLUST and local modes in the case of IMM. In addition, we attempted to fine-tune parameters in IMM to improve its convergence properties, and our preliminary results suggested that using fine-tuned parameters improved the resulting z-scores. However, due to the high computational cost (100 randomly selected subsets for each number of experiments for each dataset), it is not feasible to perform the complete convergence-tuning analysis for the IMM procedure for each dataset. In the practical setting in which parameter tuning is feasible for the dataset of interest, we expect MCLUST and IMM to have comparable performance on average.

Standardization

Standardization means that the average expression value of each gene across all experiments is subtracted from the expression value of each gene and then divided by the standard deviation of its expression levels across all experiments. It can be shown that correlation and Euclidean distance are equivalent after standardization. We showed that standardization greatly improves the z-scores from model-based clustering methods.

Additional Results

We applied different clustering algorithms (including hierarchical average-link and complete-link, model-based MCLUST and IMM) to gene subsets using *all* experiments from each of the two gene subsets of the compendium and environmental stress data.

1. Compendium data subset with 215 genes and 273 experiments

Figure A.1.a: Comparing the heuristic-based methods (hierarchical complete-link and average-link, using different similarity measures) using SCPD as the evaluation criterion on the compendium data subset with 215 genes and 273 experiments. The similarity measures we compared include correlation, Euclidean distance, error-weighted correlation and error-weighed Euclidean distance). Since the p-values of log ratios are available on the compendium data, we can compute the error-weighted similarity measures.

We showed that hierarchical complete-link using correlation produced the highest z-scores among the heuristic-based hierarchical algorithms tested.

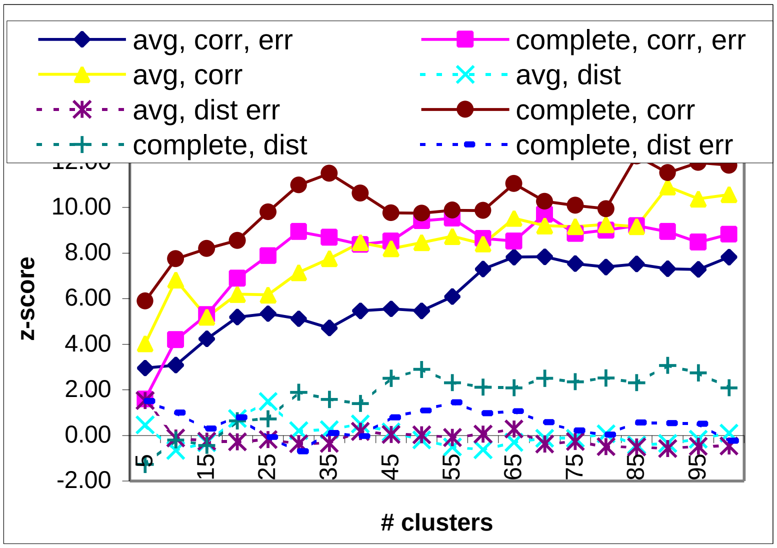


Figure A.1.b: Comparing different models in MCLUST using SCPD as the evaluation criterion on the compendium data subset with 215 genes and 273 experiments.
EII = equal-volume spherical model
VII = unequal-volume spherical model
VVV = unconstrained model
Std = standardized data
Hie = results after the model-based hierarchical step
Em = results after the EM step

We showed that:

- standardization greatly improves the z-scores
- the equal-volume spherical model (EII) produces the highest z-scores
- The EM step after the hierarchical initialization step does not quite improve the z-scores.

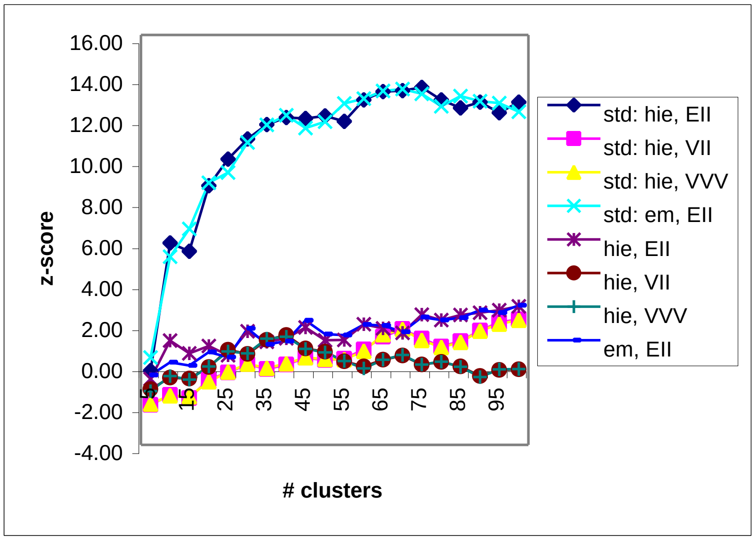
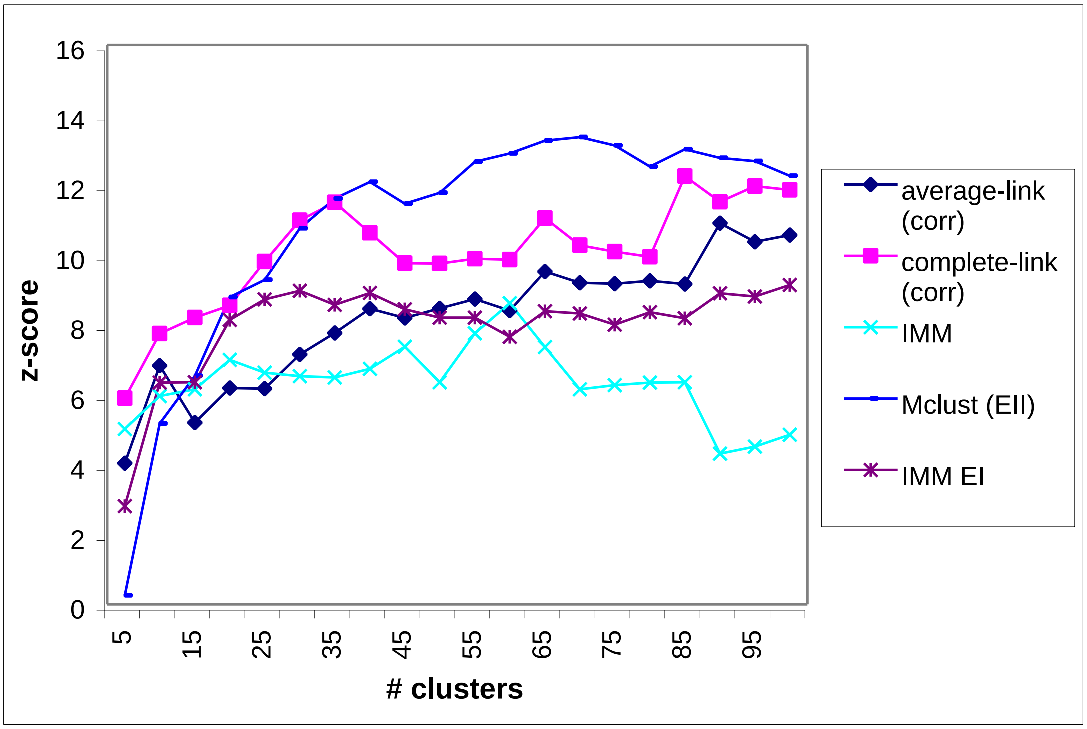


Figure A.1.c: Comparing different clustering algorithms (hierarchical complete-link and average-link with correlation, IMM with the equal and unequal volume model, MCLUST) using SCPD as the evaluation criterion on the compendium data subset with 215 genes and 273 experiments. “IMM EI” denotes results generated using IMM equal-volume spherical model.

We showed that MCLUST (with the equal volume spherical model) on the standardized data produced the highest z-scores.



2. Compendium data subset with 537 genes and 258 experiments

Figure A.2.a: Comparing the heuristic-based methods (hierarchical complete-link and average-link, using different similarity measures) using YPD as the evaluation criterion on the compendium data subset with 537 genes and 258 experiments. The similarity measures we compared include correlation, Euclidean distance, error-weighted correlation and error-weighed Euclidean distance). Since the p-values of log ratios are available on the compendium data, we can compute the error-weighted similarity measures.

We showed that hierarchical complete-link using correlation produced the highest z-scores among the heuristic-based hierarchical algorithms tested.

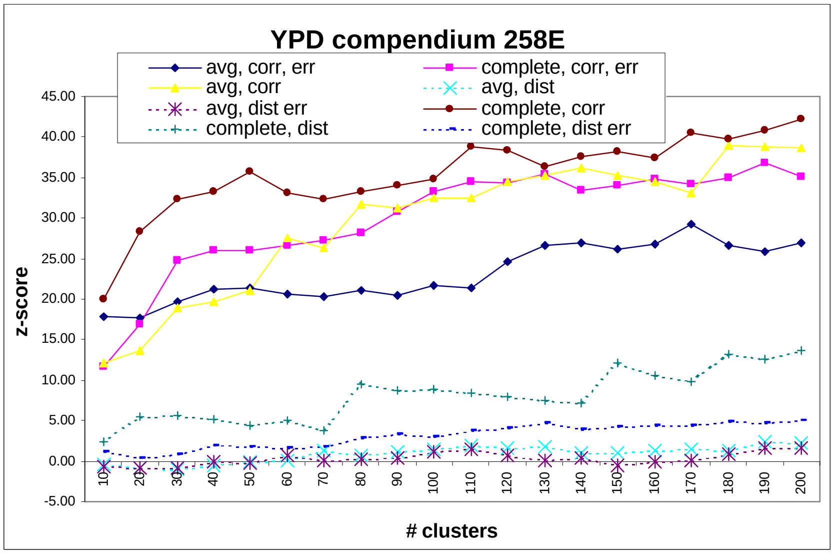
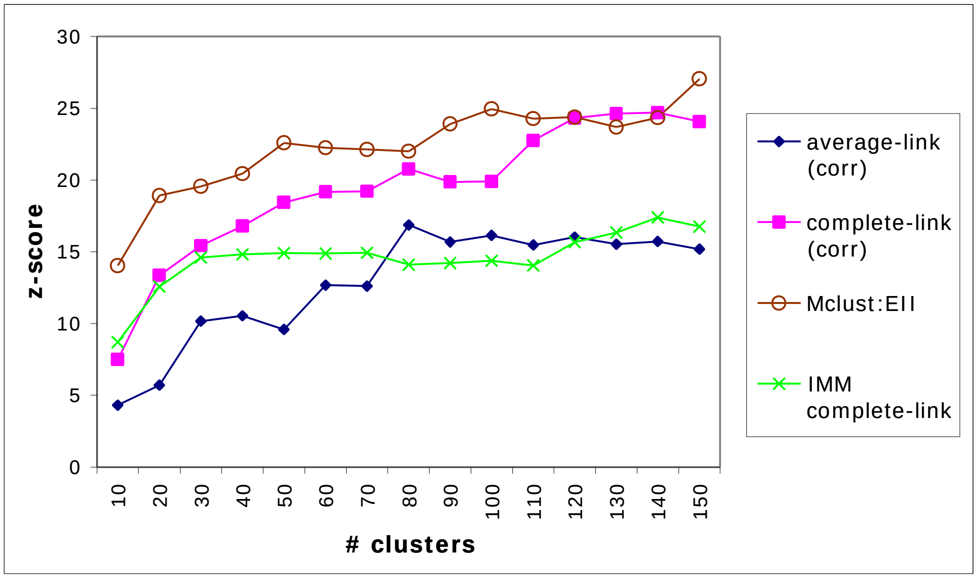


Figure A.2.b: Comparing different clustering algorithms (hierarchical complete-link and average-link with correlation, IMM, MCLUST) using ChIP data as the evaluation criterion on the compendium data subset with 537 genes and 258 experiments. “IMM” denotes results generated using IMM unequal-volume spherical model.

We showed that MCLUST (with the equal volume spherical model) on the standardized data produced the highest z-scores.



3. Environmental stress data subset with 205 genes and 205 experiments

Figure A.3.a: Comparing different clustering algorithms (hierarchical complete-link and average-link with correlation, IMM, MCLUST) using SCPD as the evaluation criterion on the environmental stress data subset with 205 genes and 205 experiments. “IMM” denotes results generated using IMM unequal-volume spherical model.

We showed that MCLUST (with the equal volume spherical model) on the standardized data produced the highest z-scores.

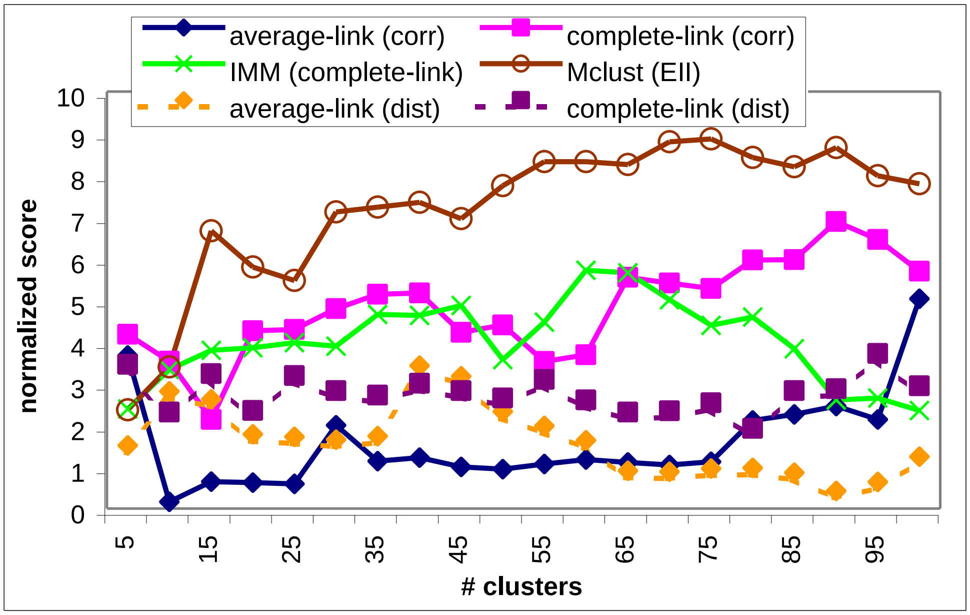
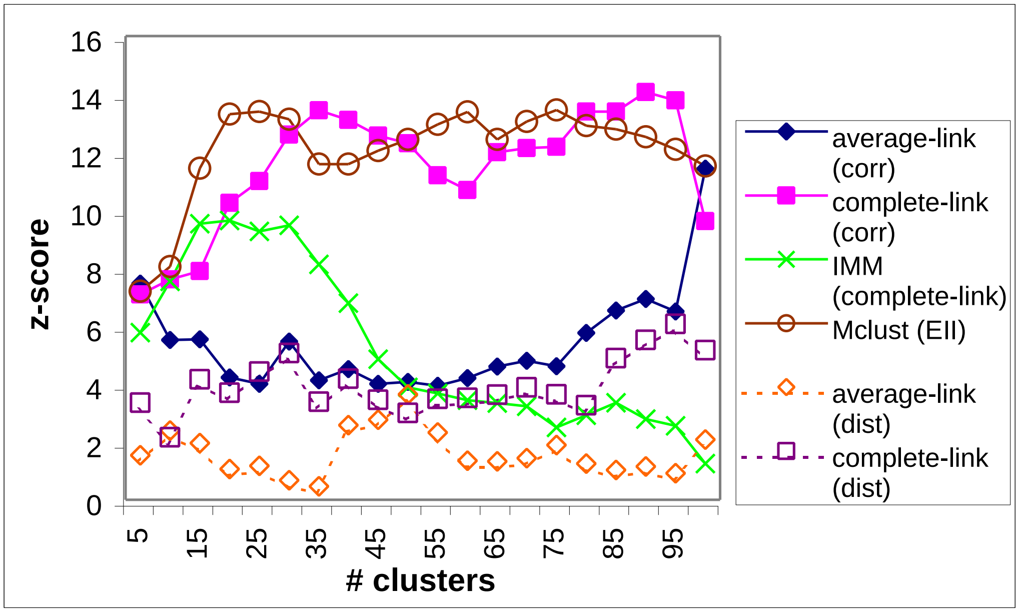


Figure A.3.b: Comparing different clustering algorithms (hierarchical complete-link and average-link with correlation, IMM, MCLUST) using ChIP data as the evaluation criterion on the environmental stress data subset with 205 genes and 205 experiments. "IMM" denotes results generated using IMM unequal-volume spherical model.

We showed that MCLUST (with the equal volume spherical model) on the standardized data and hierarchical complete-link using correlation produced the highest z-scores.



4. Environmental stress data subset with 526 genes and 198 experiments

Figure A.4.a: Comparing different clustering algorithms (hierarchical complete-link and average-link with correlation, IMM, MCLUST) using YPD as the evaluation criterion on the environmental stress data subset with 526 genes and 198 experiments. "IMM" denotes results generated using IMM unequal-volume spherical model.

We showed that MCLUST (with the equal volume spherical model) on the standardized data produced the highest z-scores.

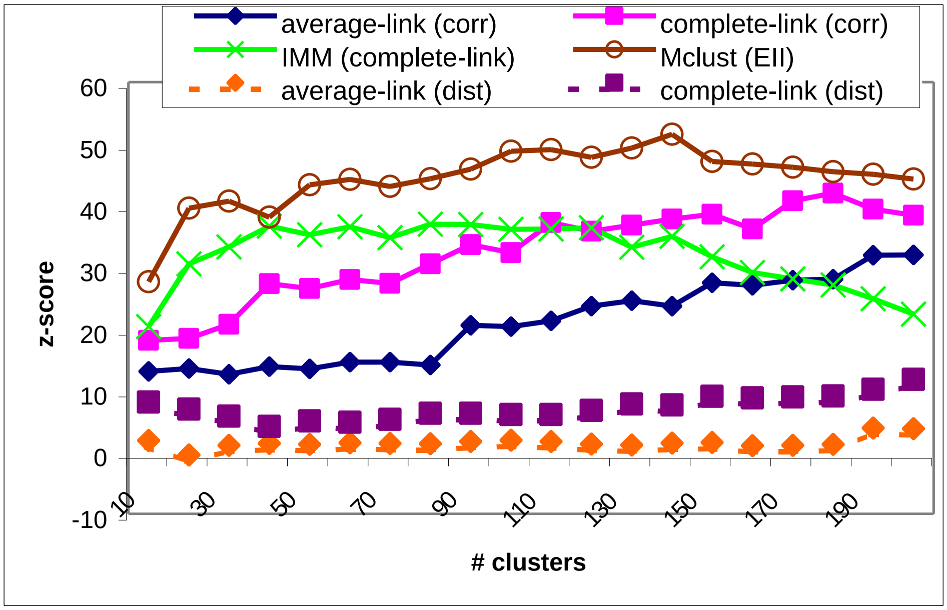
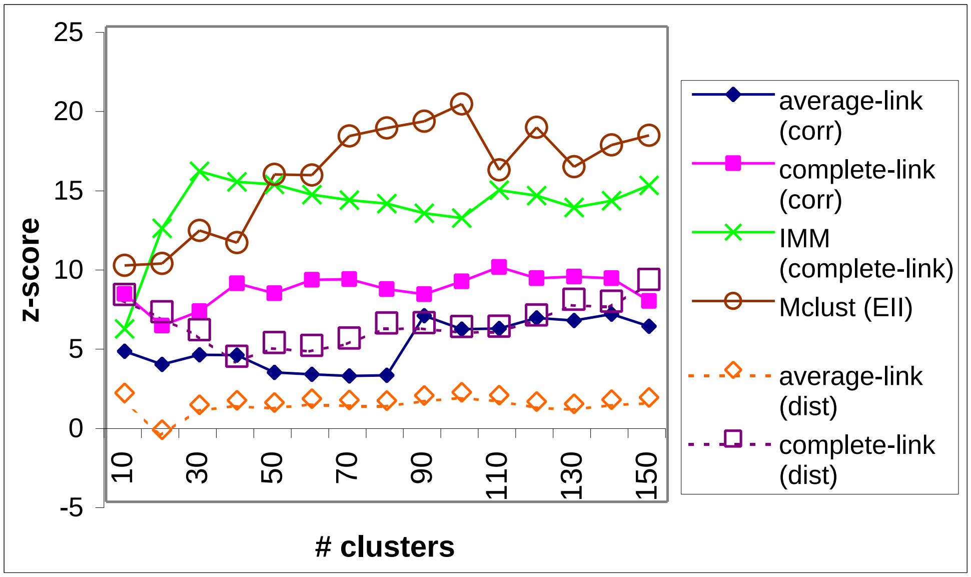


Figure A.4.b: Comparing different clustering algorithms (hierarchical complete-link and average-link with correlation, IMM, MCLUST) using ChIP data as the evaluation criterion on the environmental stress data subset with 526 genes and 198 experiments. “IMM” denotes results generated using IMM unequal-volume spherical model.

We showed that MCLUST (with the equal volume spherical model) and IMM on the standardized data produced the highest z-scores.



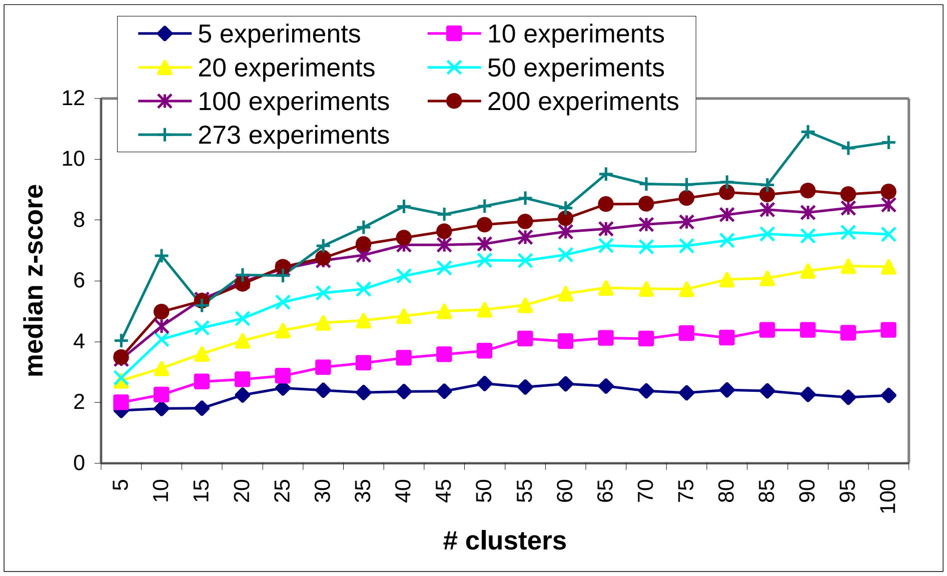
B. Effect of number of microarray experiments

We applied different clustering algorithms (including hierarchical average-link and complete-link, model-based MCLUST and IMM) to gene subsets using *different numbers* of microarray experiments. In order to produce typical datasets with E experiments (where E = 5, 10, 20, 50, 100), we randomly sampled (with replacement) 100 different subsets of E experiments from the compendium data with 215 genes and 273 experiments. The ability to identify co-regulated genes from clustering results is summarized by the median z-scores over the 100 randomly sampled datasets. A high median z-score indicates a high proportion of co-regulated genes from clustering results compared to those from random partitions.

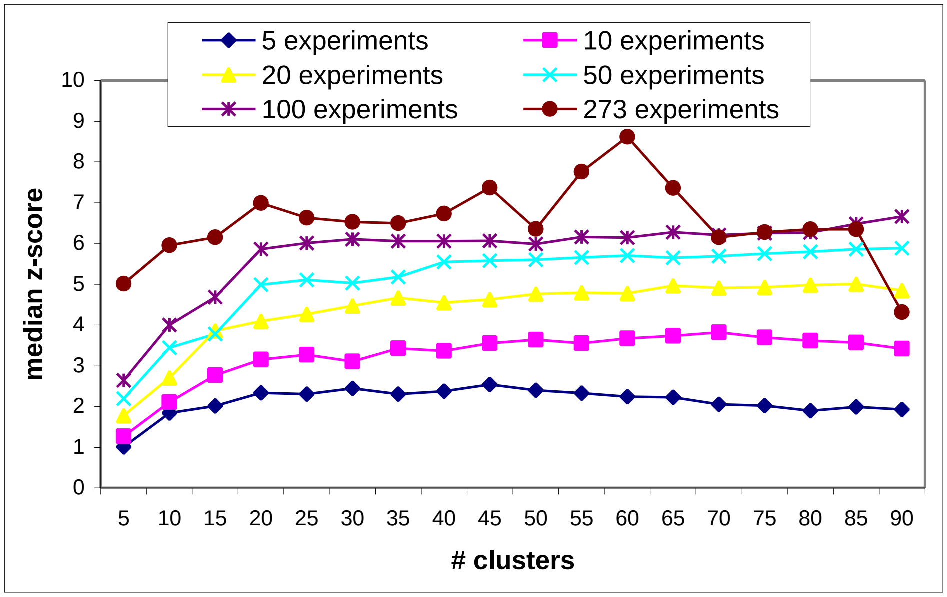
1. Compendium data subset with 215 genes and 273 experiments

Figures B.1.a to Figures B.1.d: Using SCPD as the evaluation criterion, we compared the median z-scores using different numbers of experiments (E) from different clustering algorithms over a range of different numbers of clusters (5 to 100) on the compendium data subset with 215 genes. The median z-scores generally increase as E increases over different numbers of clusters. This shows that higher proportions of co-regulated genes are identified on microarray datasets with higher numbers of experiments.

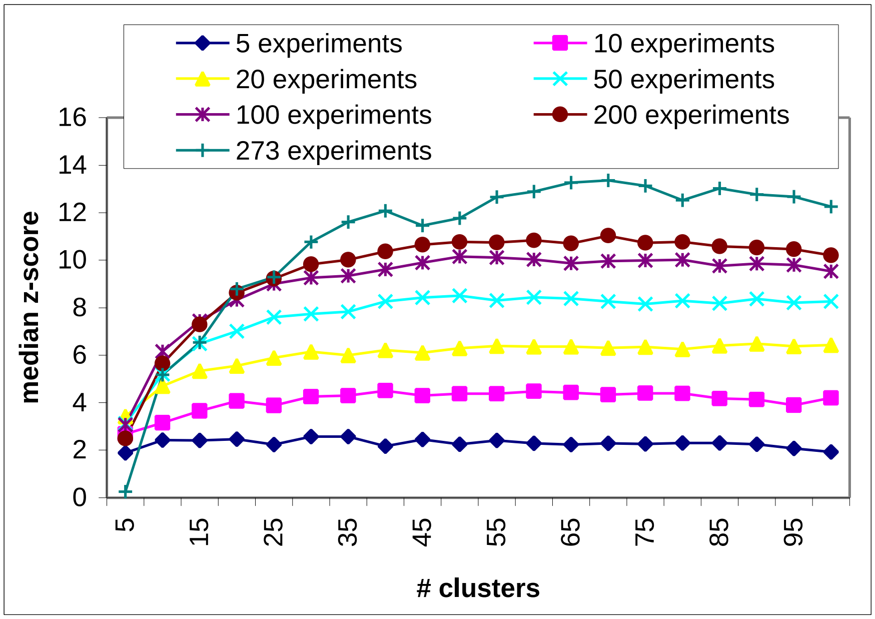
Figures B.1.a: Comparing the median z-scores using hierarchical average-link and correlation over different E.



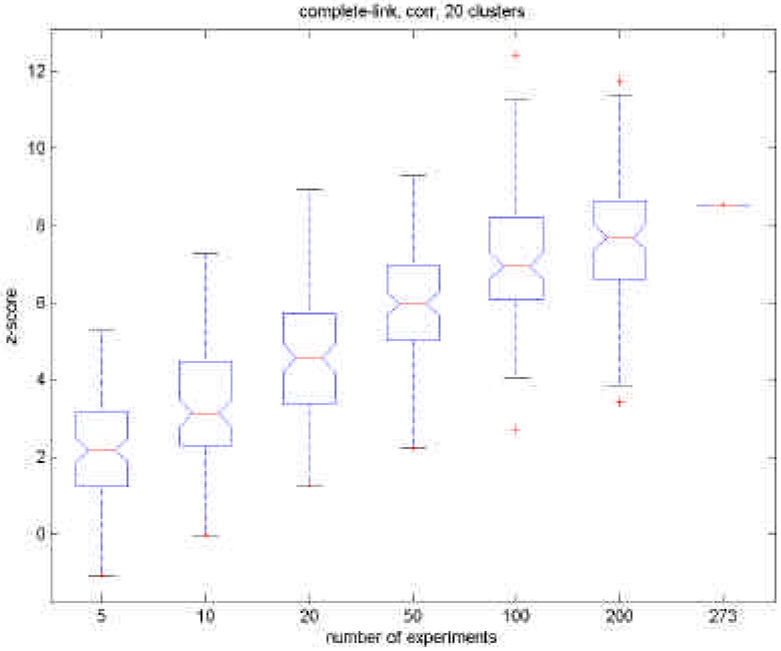
Figures B.1.b: Comparing the median z-scores using IMM (unequal volume spherical model) over different E.



Figures B.1.c: Comparing the median z-scores using MCLUST (equal volume spherical model) over different E.

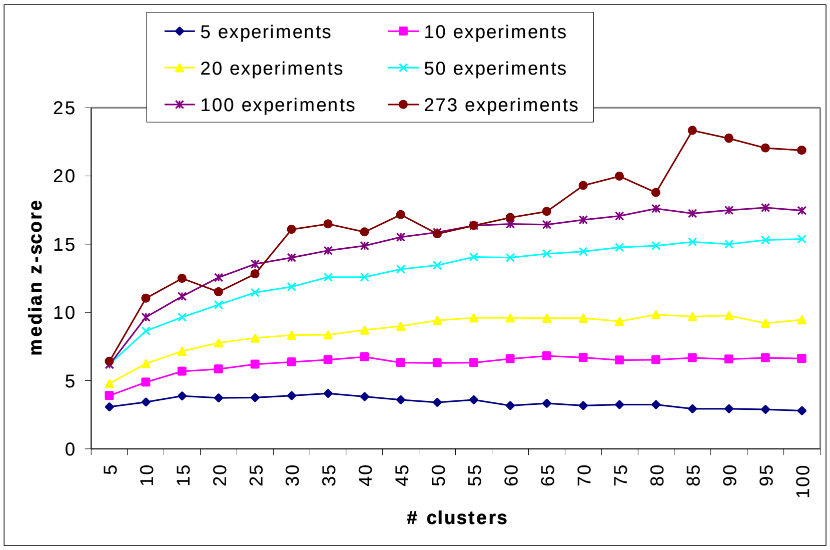


Figures B.1.d: Comparing the distribution of z-scores using hierarchical average-link using correlation over different E at 20 clusters.

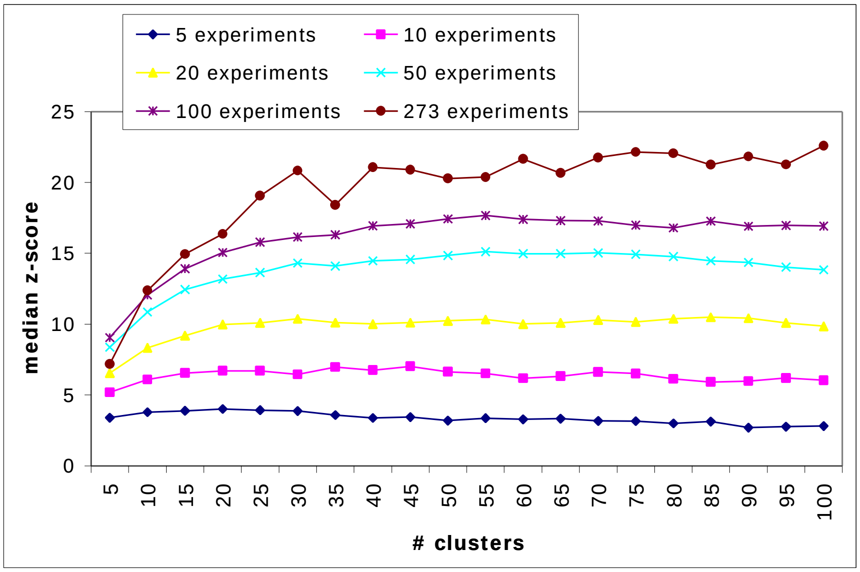


Figures B.1.e to Figures B.1.f: Using ChIP data as our evaluation criterion, we compared the median z-scores using different numbers of experiments (E) from different clustering algorithms over a range of different numbers of clusters (5 to 100) on the compendium data subset with 215 genesThe median z-scores generally increase as E increases over different numbers of clusters. This shows that higher proportions of co-regulated genes are identified on microarray datasets with higher numbers of experiments.

Figures B.1.e: Comparing the median z-scores using hierarchical complete-link and correlation over different E.



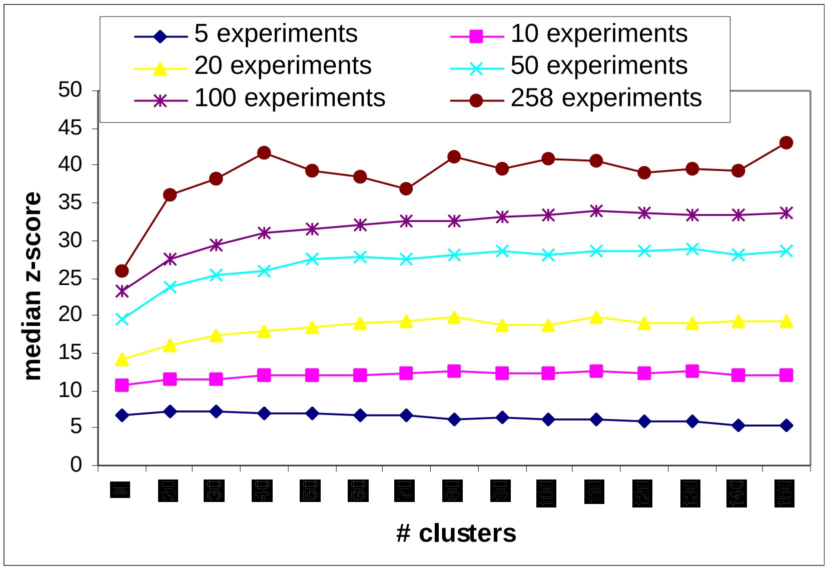
Figures B.1.f: Comparing the median z-scores using MCLUST (equal volume spherical model) over different E.



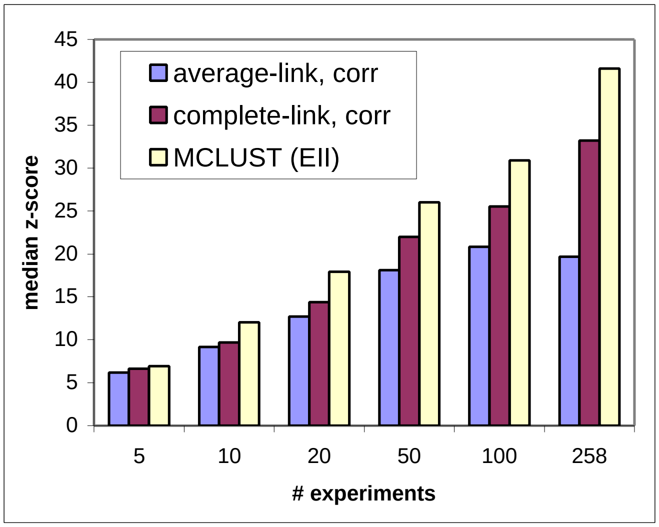
2. Compendium data subset with 537 genes and 258 experiments

Figures B.2.a to Figures B.2.b: Using YPD as the evaluation criterion, we compared the median z-scores using different numbers of experiments (E) from different clustering algorithms over a range of different numbers of clusters (5 to 100) on the compendium data subset with 537 genes. The median z-scores generally increase as E increases over different numbers of clusters. This shows that higher proportions of co-regulated genes are identified on microarray datasets with higher numbers of experiments.

Figures B.2.a: Comparing the median z-scores using MCLUST (equal volume spherical model) over different E.

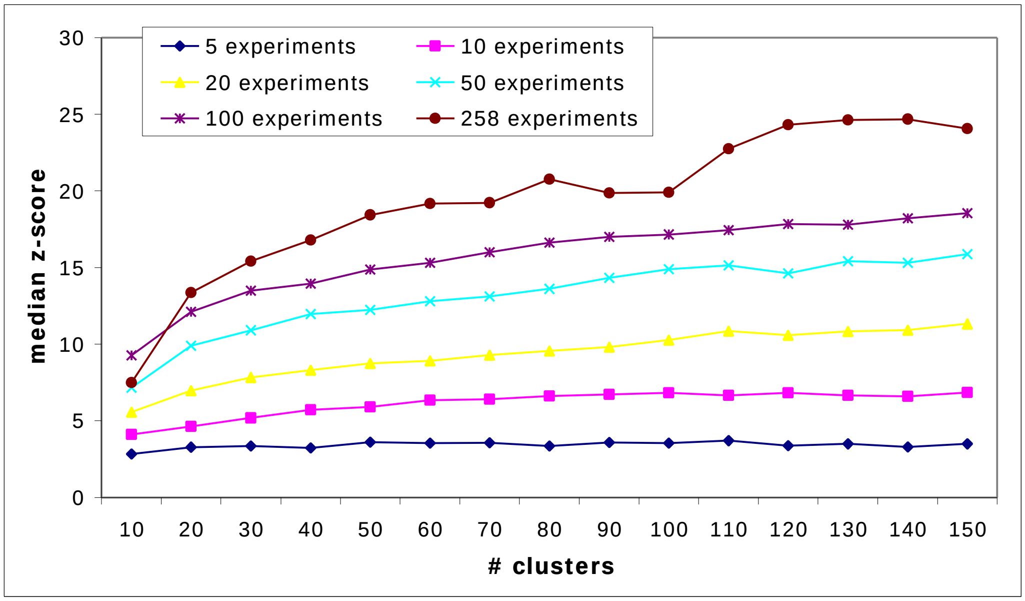


Figures B.2.b: Comparing the median z-scores using different clustering algorithms over different E at 40 clusters.

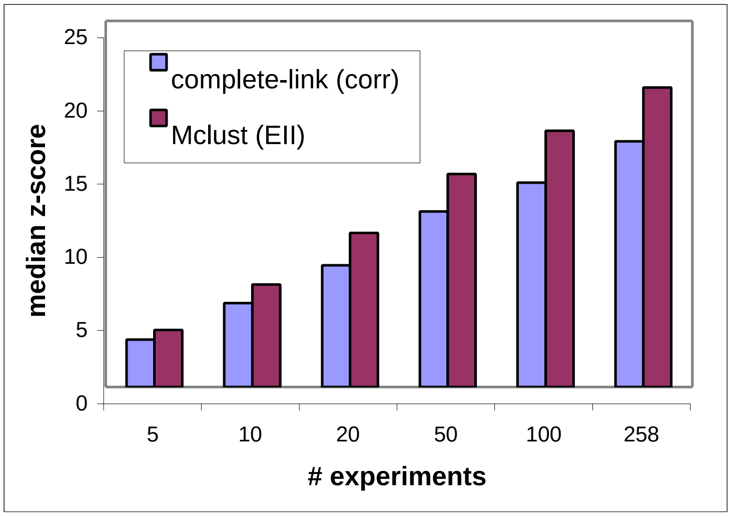


Figures B.2.c to Figures B.2.d: Using ChIP data as the evaluation criterion, we compared the median z-scores using different numbers of experiments (E) from different clustering algorithms over a range of different numbers of clusters (5 to 100) on the compendium data subset with 537 genes. The median z-scores generally increase as E increases over different numbers of clusters. This shows that higher proportions of co-regulated genes are identified on microarray datasets with higher numbers of experiments.

Figures B.2.c: Comparing the median z-scores using complete-link and correlation over different E.



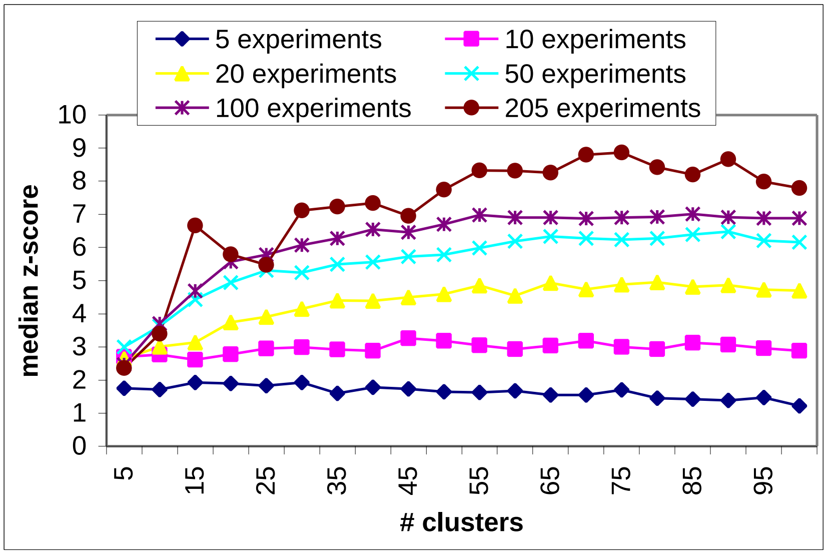
Figures B.2.d: Comparing the median z-scores using different clustering algorithms over different E at 40 clusters.



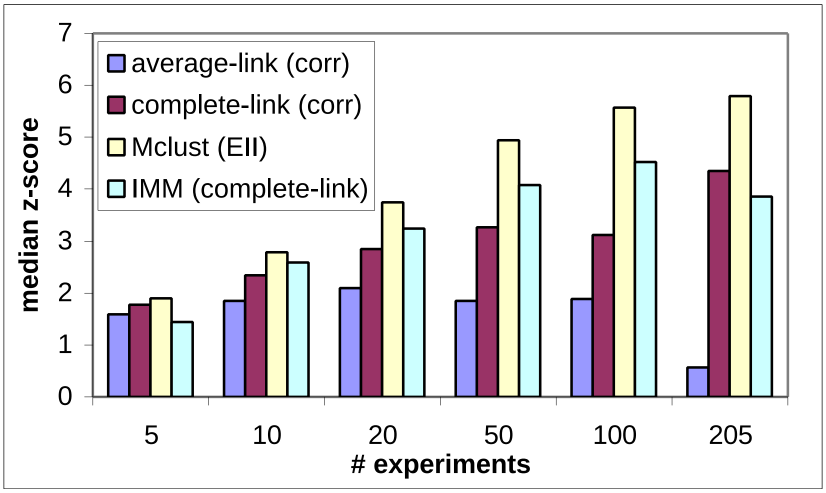
3. Environmental stress data subset with 205 genes and 205 experiments

Figures B.3.a to Figures B.3.b: Using SCPD as the evaluation criterion, we compared the median z-scores using different numbers of experiments (E) from different clustering algorithms over a range of different numbers of clusters (5 to 100) on the environmental stress data subset with 205 genes. The median z-scores generally increase as E increases over different numbers of clusters. This shows that higher proportions of co-regulated genes are identified on microarray datasets with higher numbers of experiments.

Figures B.3.a: Comparing the median z-scores using MCLUST (equal volume spherical model) over different E.

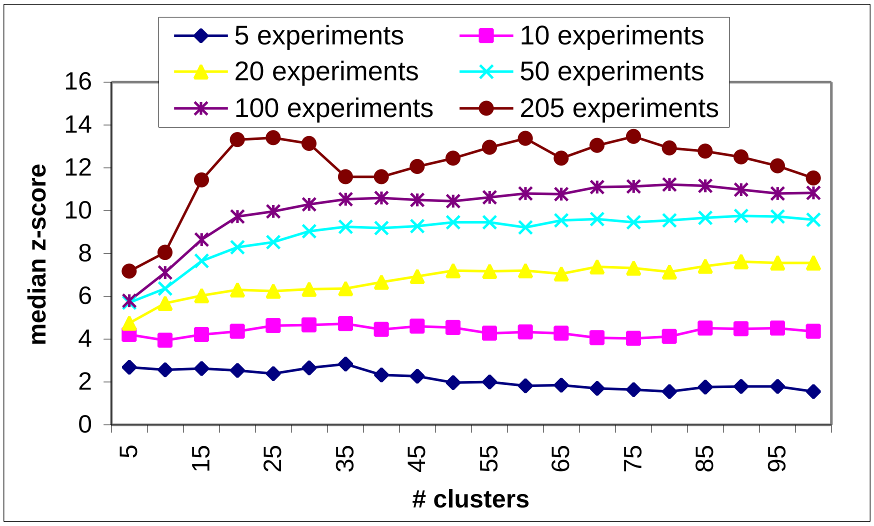


Figures B.3.b: Comparing the median z-scores using different clustering algorithms over different E at 20 clusters.

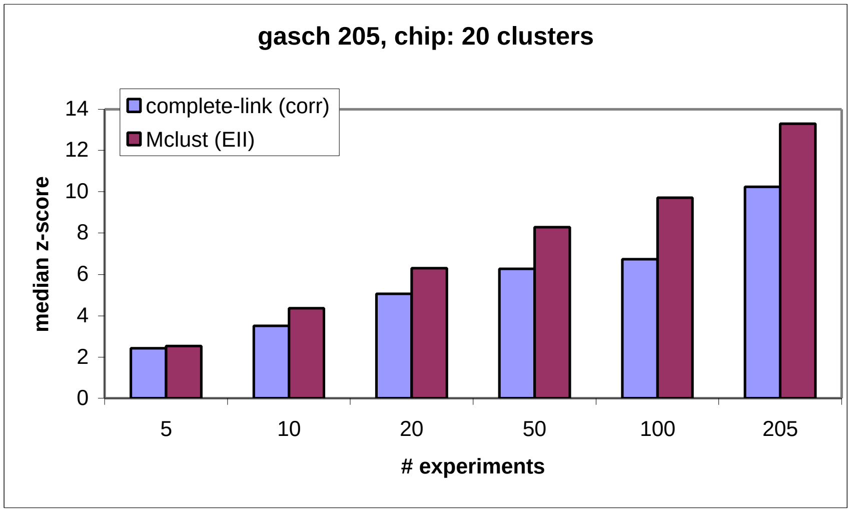


Figures B.3.c to Figures B.3.d: Using ChIP data as the evaluation criterion, we compared the median z-scores using different numbers of experiments (E) from different clustering algorithms over a range of different numbers of clusters (5 to 100) on the environmental stress data subset with 205 genes. The median z-scores generally increase as E increases over different numbers of clusters. This shows that higher proportions of co-regulated genes are identified on microarray datasets with higher numbers of experiments.

Figures B.3.c: Comparing the median z-scores using MCLUST (equal volume spherical model) over different E.



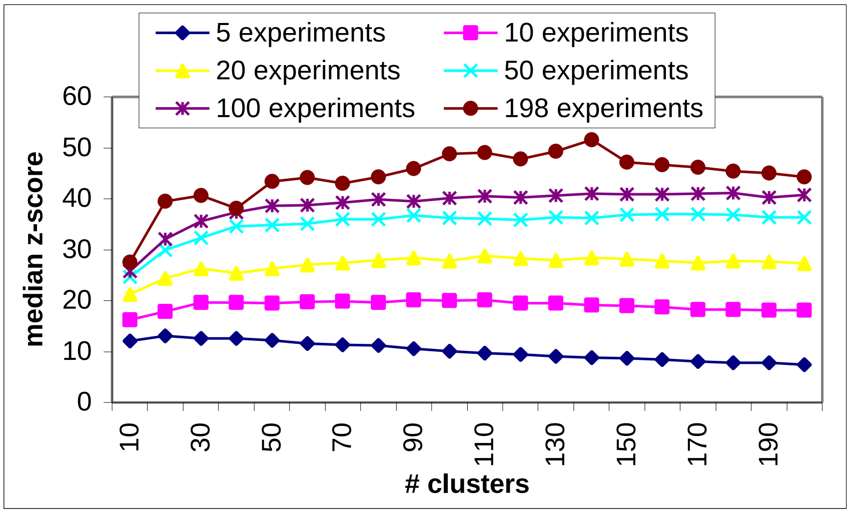
Figures B.3.d: Comparing the median z-scores using different clustering algorithms over different E at 20 clusters.



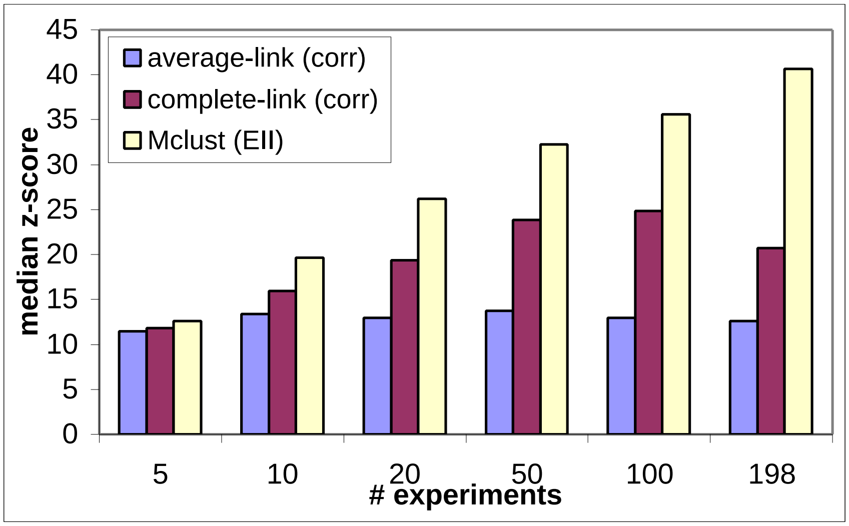
4. Environmental stress data subset with 526 genes and 198 experiments

Figures B.4.a to Figures B.4.b: Using YPD as the evaluation criterion, we compared the median z-scores using different numbers of experiments (E) from different clustering algorithms over a range of different numbers of clusters (5 to 100) on the environmental stress data subset with 526 genes. The median z-scores generally increase as E increases over different numbers of clusters. This shows that higher proportions of co-regulated genes are identified on microarray datasets with higher numbers of experiments.

Figures B.4.a: Comparing the median z-scores using MCLUST (equal volume spherical model) over different E.

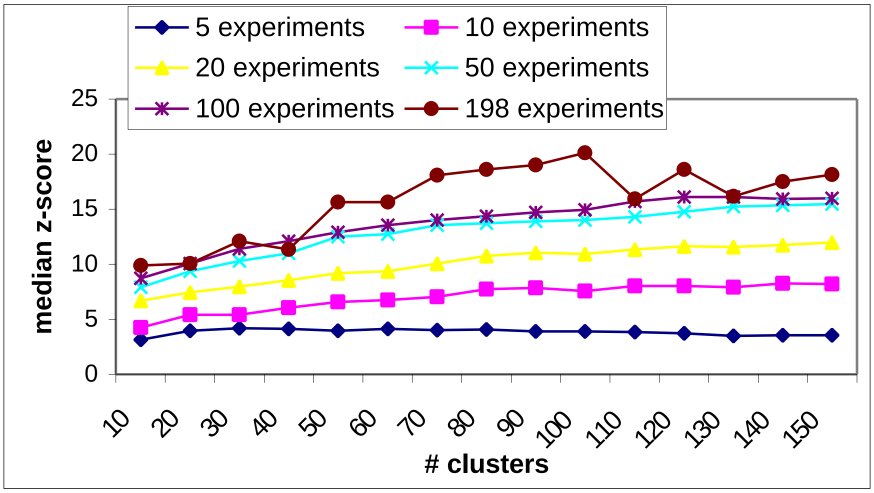


Figures B.4.b: Comparing the median z-scores using different clustering algorithms over different E at 30 clusters.

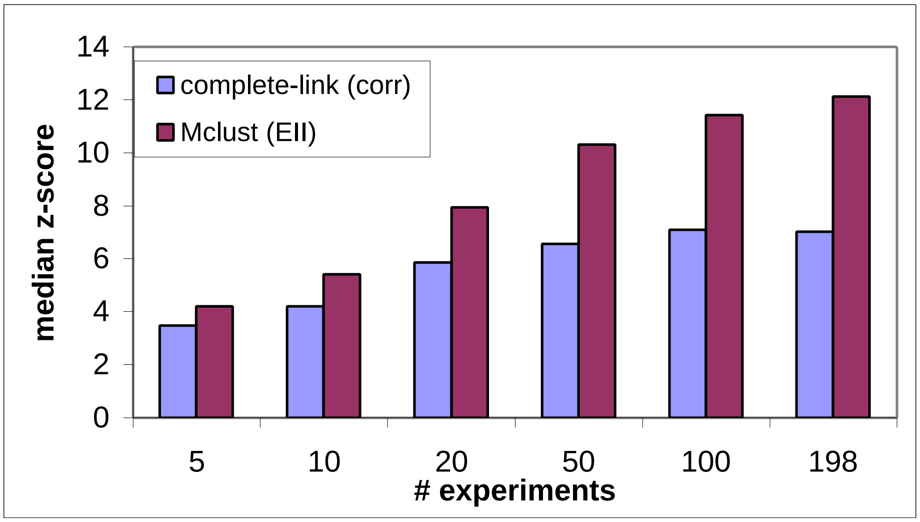


Figures B.4.c to Figures B.4.d: Using ChIP data as the evaluation criterion, we compared the median z-scores using different numbers of experiments (E) from different clustering algorithms over a range of different numbers of clusters (5 to 100) on the environmental stress data subset with 526 genes. The median z-scores generally increase as E increases over different numbers of clusters. This shows that higher proportions of co-regulated genes are identified on microarray datasets with higher numbers of experiments.

Figures B.4.c: Comparing the median z-scores using MCLUST (equal volume spherical model) over different E.



Figures B.4.d: Comparing the median z-scores using different clustering algorithms over different E at 30 clusters.



5. Compendium data subset with **901 genes** and 252 experiments

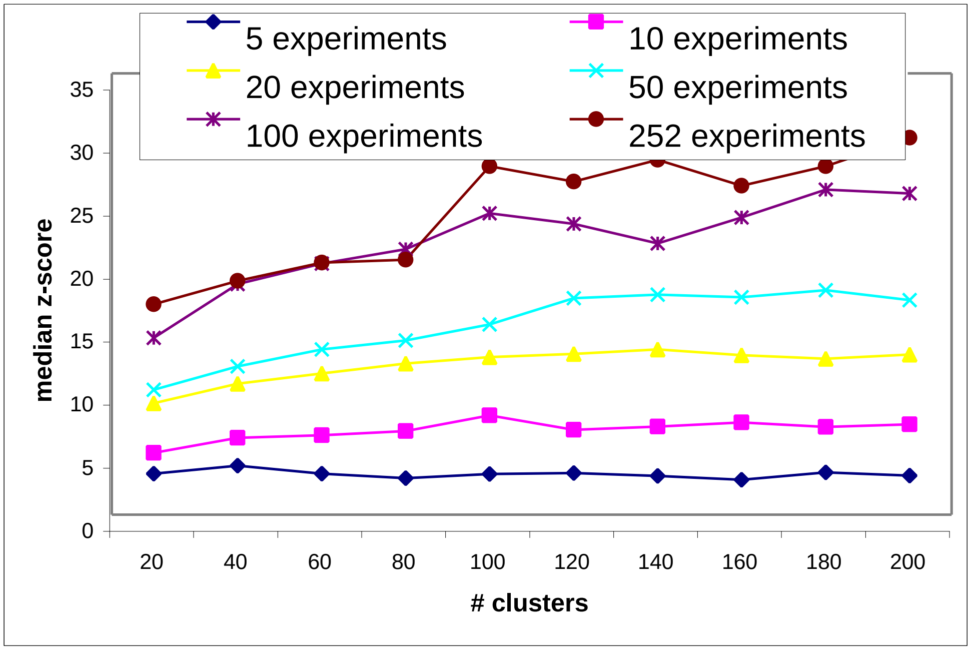
Goal: Results in section B.5 aims to provide evidence for our general results when ***larger data subsets*** are used in our study.

Rationale: The gene subsets used in sections B.1 (compendium data susbset with 215 genes) and B.3 (environmental stress data subset with 205 genes) are chosen based on the genes listed in SCPD, while the gene subsets used in sections B.2 (compendium data susbset with 537 genes) and B.4 (environmental stress data subset with 526 genes) are chosen based on the genes listed in YPD. However, the ChIP technology is a global survey of the interactions between genes and specific transcription factors. A total of 2343 genes are bound to at least one transcription factors in the ChIP data (Lee et al. 2002). Initially, we tried to study the effect of different numbers of experiments on this set of 2343 genes on the compendium data. However, the computational running time is extremely intensive. Therefore, we randomly selected approximately 1000 genes from this 2343 gene list, extracted the subset of compendium dataset from these 1000 genes, and removed the genes and experiments with many missing values. This procedure results in 901 genes and 252 experiments. Due to the computational running time, we randomly selected n=10 subsets for each E, and generated N=200 random partitions to compute the z-score.

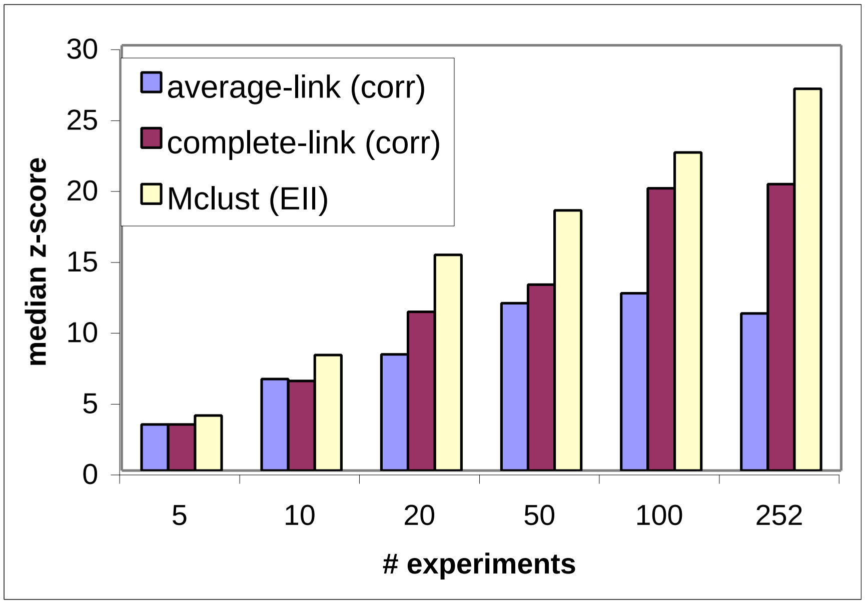
Observation: This larger subset shows the same general results as the smaller gene subsets.

Figures B.5.a to B.5.b: Using ChIP data as the evaluation criterion, we compared the median z-scores using different numbers of experiments (E) from different clustering algorithms over a range of different numbers of clusters (20 to 200) on the compendium data subset with 901 genes. The median z-scores generally increase as E increases over different numbers of clusters. This shows that higher proportions of co-regulated genes are identified on microarray datasets with higher numbers of experiments.

Figures B.5.a: Comparing the median z-scores using complete-link and correlation over different E.



Figures B.5.b: Comparing the median z-scores using different clustering algorithms over different E at 60 clusters. The optimal number of clusters is estimated to be 66, using auto-IMM.



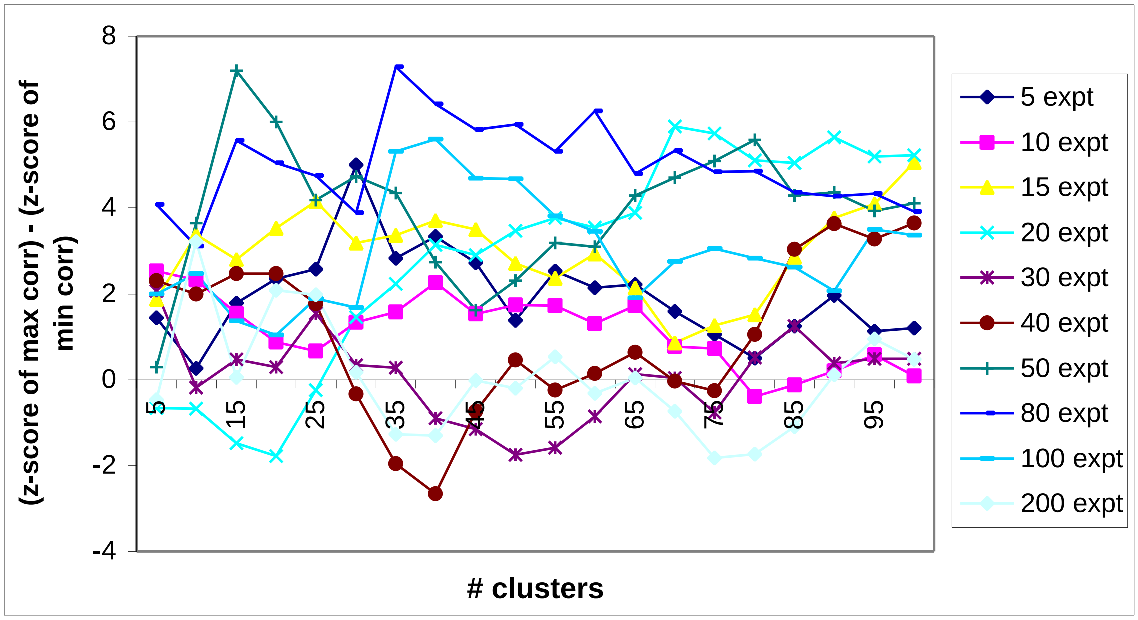
C. Effect of diversity of experiments

In Figures C.1 to C.4, the difference between the z-score of a data subset with low diversity and the z-score of a data subset with high diversity is plotted against the number of clusters. The data subsets with high and low diversity are obtained using a greedy algorithm described in the Methods section in the manuscript. A positive difference (y-axis) implies that microarray experiments that are similar (hence, low diversity) tend to produce clusters with relatively high proportions of co-regulated genes. On the other hand, a negative difference (y-axis) implies that experiments that are diverse tend to produce clusters with high proportions of co-regulated genes.

1. Compendium data subset with 215 genes and 273 experiments

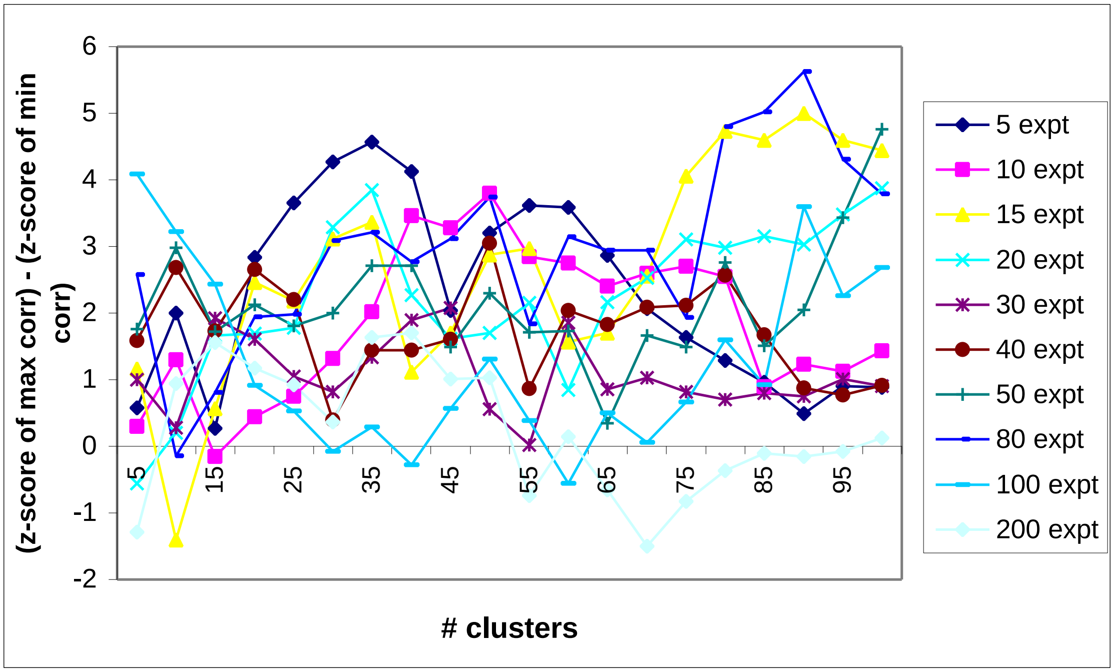
Figures C.1.a: Comparing the difference in z-scores using hierarchical complete-link and correlation, evaluated using SCPD.

The results show that experiments that are similar tend to produce higher z-scores. However, there are a few exceptions (e.g. E = 30, 40, 200).



Figures C.1.b: Comparing the difference in z-scores using hierarchical average-link and correlation, evaluated using SCPD.

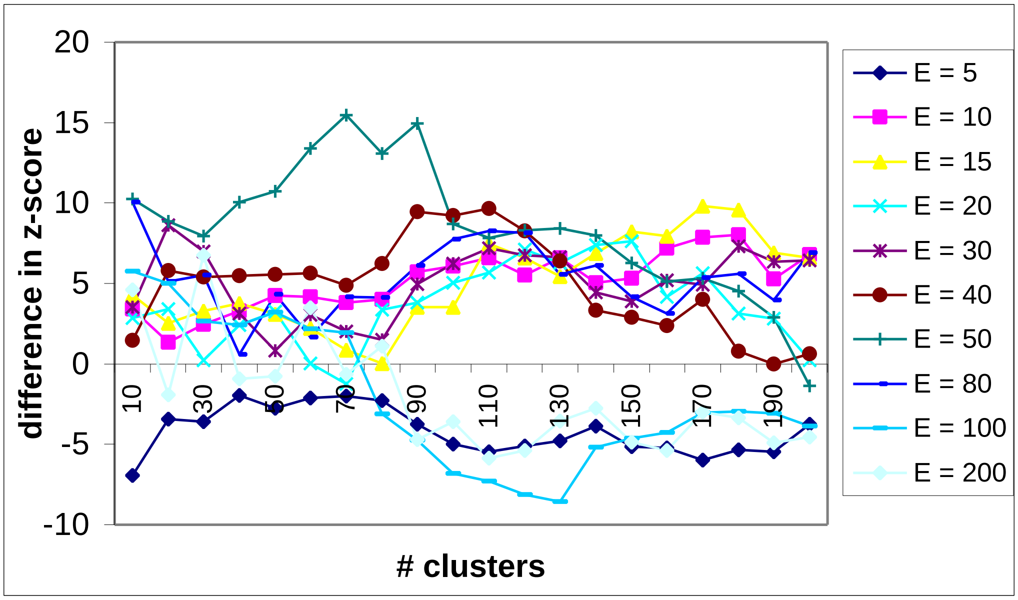
The results show that experiments that are similar tend to produce higher z-scores. However, there are a few exceptions (e.g. E = 15).


(z-score of max corr) - (z-score of min corr)
6
5
4
3
2
1
0
-1
-2
5
15
25
35
45
55
65
75
85
95
# clusters
5 expt
10 expt
15 expt
20 expt
30 expt
40 expt
50 expt
80 expt
100 expt
200 expt

2. Compendium data subset with 537 genes and 258 experiments

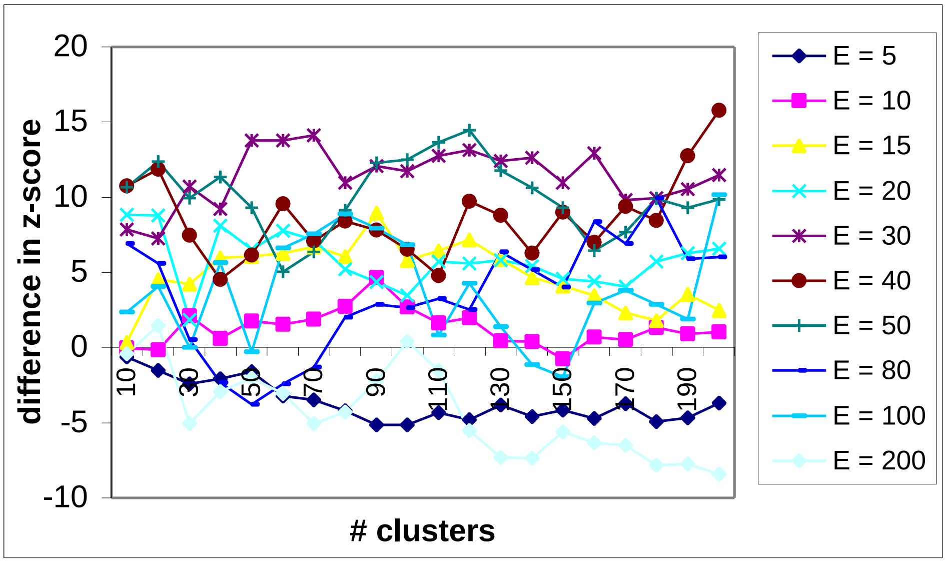
Figures C.2.a: Comparing the difference in z-scores using hierarchical average-link and correlation, evaluated using YPD.

The results show that experiments that are similar tend to produce higher z-scores. However, there are a few exceptions (e.g. E = 5, 100, 200).



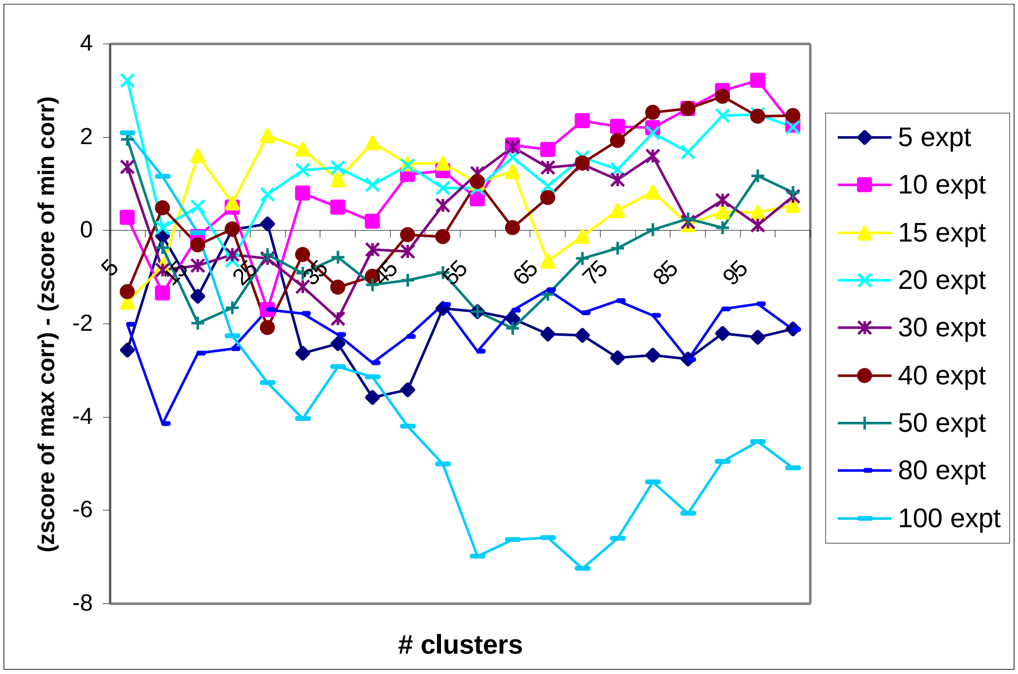
Figures C.2.b: Comparing the difference in z-scores using hierarchical complete-link and correlation, evaluated using YPD.

The results show that experiments that are similar tend to produce higher z-scores. However, there are a few exceptions (e.g. E = 5, 100, 200).

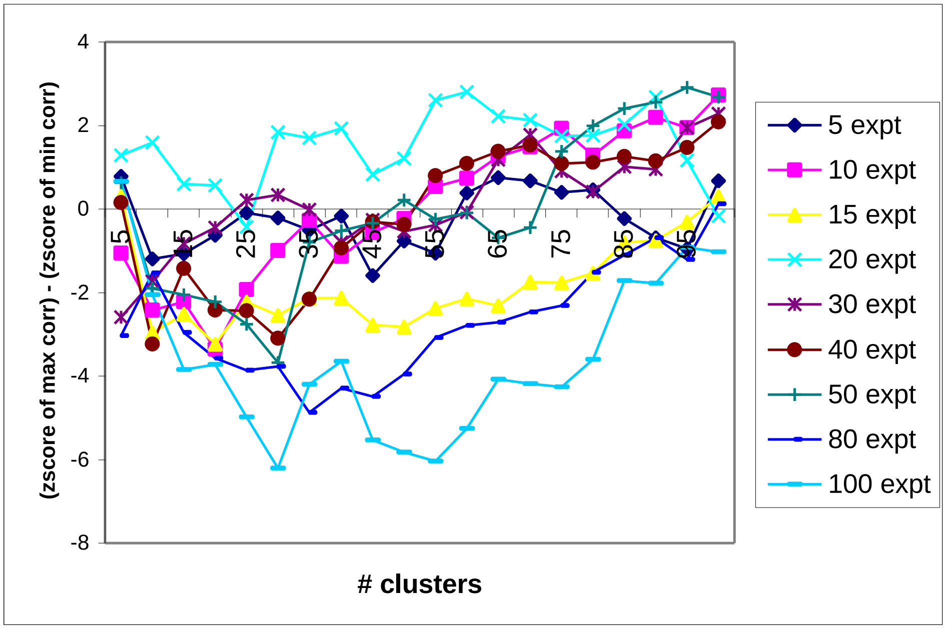


3. Environmental stress data subset with 205 genes and 205 experiments

Figures C.3.a: Comparing the difference in z-scores using hierarchical complete-link and correlation, evaluated using SCPD.
There are no clear trends.

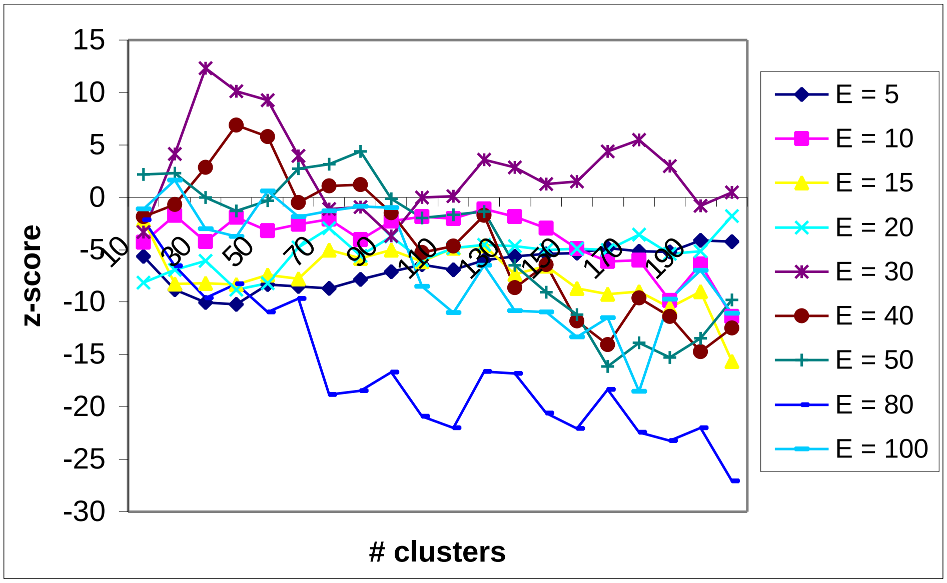


Figures C.3.b: Comparing the difference in z-scores using hierarchical average-link and correlation, evaluated using SCPD.
There are no clear trends.

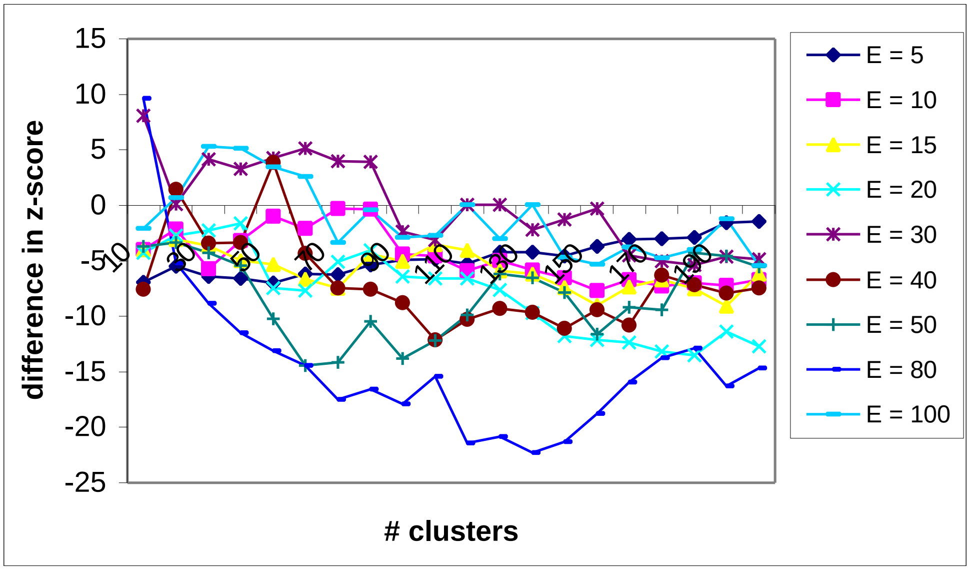


4. Environmental stress data subset with 526 genes and 198 experiments

Figures C.4.a: Comparing the difference in z-scores using hierarchical average-link and correlation, evaluated using YPD.
Experiments with high diversity tend to produce higher z-scores.



Figures C.4.b: Comparing the difference in z-scores using hierarchical complete-link and correlation, evaluated using YPD.
Experiments with high diversity tend to produce higher z-scores.

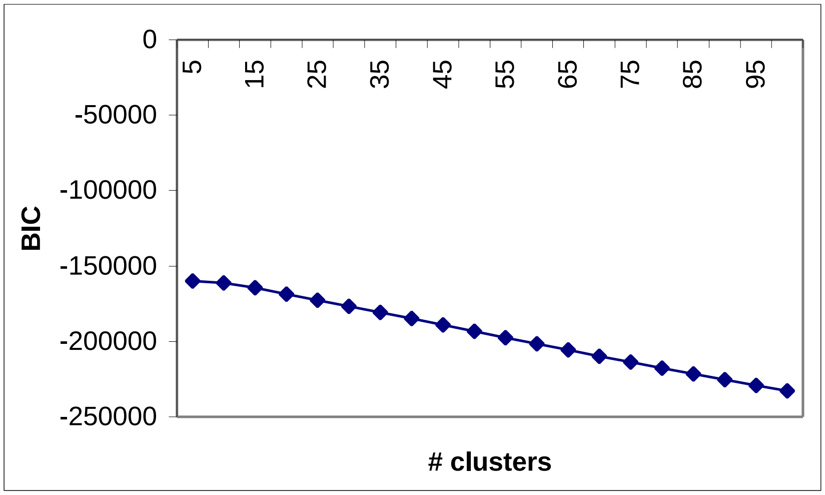


D. Estimating the optimal numbers of clusters

The major advantage of the model-based approach over traditional heuristic-based methods is that the probability framework allows us to estimate the optimal number of clusters that fit the data. In MCLUST, the Bayesian Information Criterion (BIC) is used to estimate the number of clusters. The general practice is to apply different models of MCLUST over a range of numbers of clusters and compute the BIC score for each clustering result, and then the optimal number of clusters and the best model is estimated to be the clustering result that yields the maximum BIC score. In our previous work (Yeung et al. 2001), we showed that MCLUST produces reasonable estimates of the number of clusters in addition to high cluster quality. However, MCLUST did not produce any reasonable estimates of the number of clusters in this work because neither a local or global maximum BIC score is observed (see Figures D.1 to D.4 from Supplementary Materials).
Note that no gold standard is used in computing BIC.

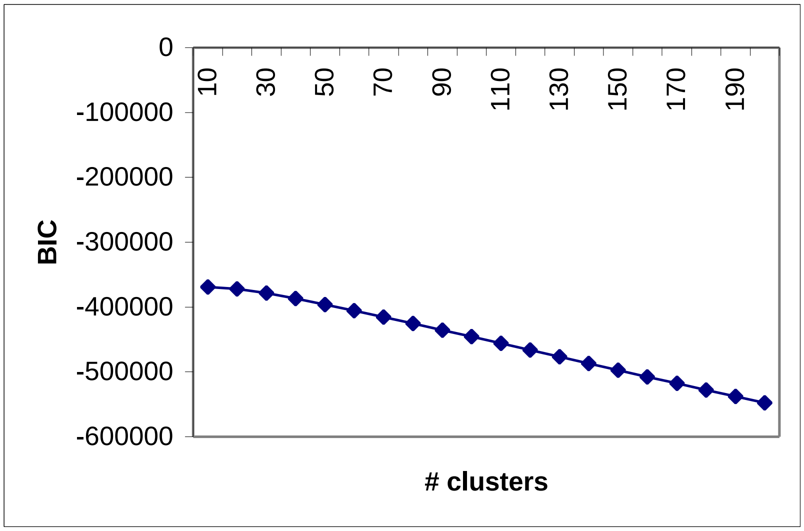
1. Compendium data subset with 215 genes and 273 experiments

Figures D.1: BIC is plotted against number of clusters (from 5 to 100) on the compendium data subset with 215 genes and 273 experiments. No local maximum is observed.



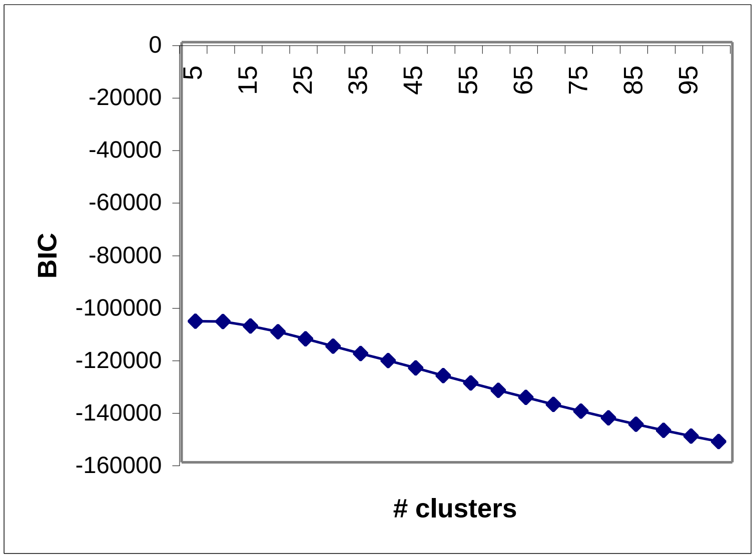
2. Compendium data subset with 537 genes and 258 experiments

Figures D.2: BIC is plotted against number of clusters (from 10 to 200) on the compendium data subset with 537 genes and 258 experiments. No local maximum is observed.



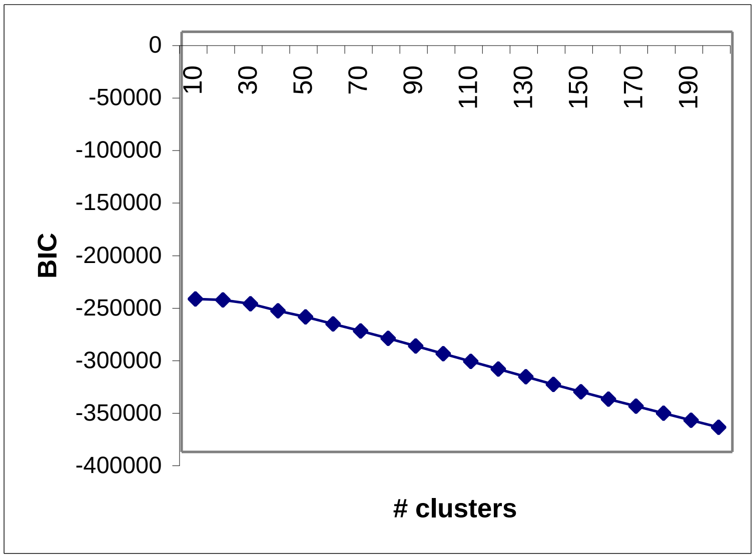
3. Environmental stress data subset with 205 genes and 205 experiments

Figures D.3: BIC is plotted against number of clusters (from 5 to 100) on the environmental stress data subset with 205 genes and 205 experiments. No local maximum is observed.



4. Environmental stress data subset with 526 genes and 198 experiments

Figures D.3: BIC is plotted against number of clusters (from 10 to 200) on the environmental stress data subset with 526 genes and 198 experiments. No local maximum is observed.



E. Results in terms of true positive (TP) rates

In this section, we show the results in terms of the TP rates, rather than z-scores shown in section B.

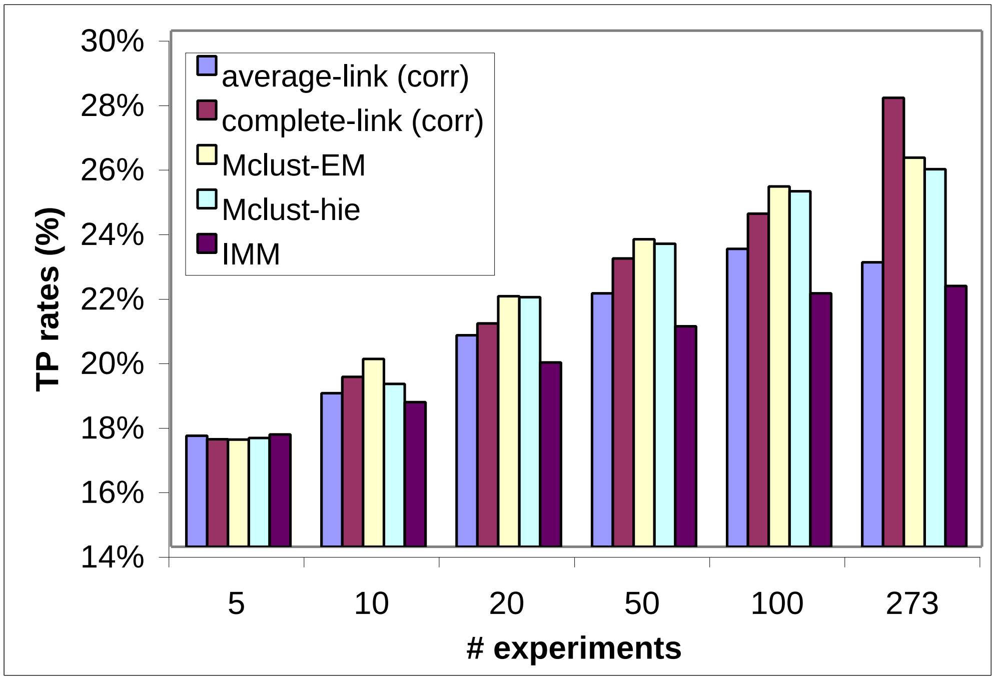
$$TP\ rate = \frac{\text{\# gene pairs from the same clusters and share at least one common transcription factor}}{\text{\# gene pairs from the same clusters}}$$

Unlike the z-scores which compare the TP rates from a given algorithm to those from random partitions, TP rate by itself is an "absolute" measure.

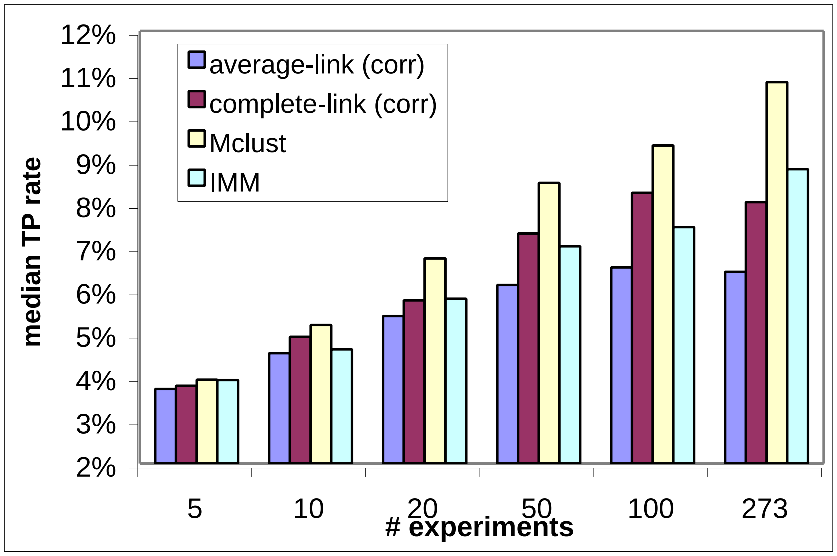
1. Compendium data subset with 215 genes and 273 experiments

Figures E.1.a: Using SCPD as the evaluation criterion, we compared the median TP rates using different numbers of experiments (E) from different clustering algorithms at *25 clusters* on the compendium data subset with 215 genes. The median TP rates generally increase as E increases over different numbers of clusters.

"Mclust-hie" represents the results from MCLUST after the hierarchical initialization step, while "Mclust-EM" represents the results from MCLUST after the EM step (i.e. final results). Our results showed that the EM step does not significantly improve the TP rates.

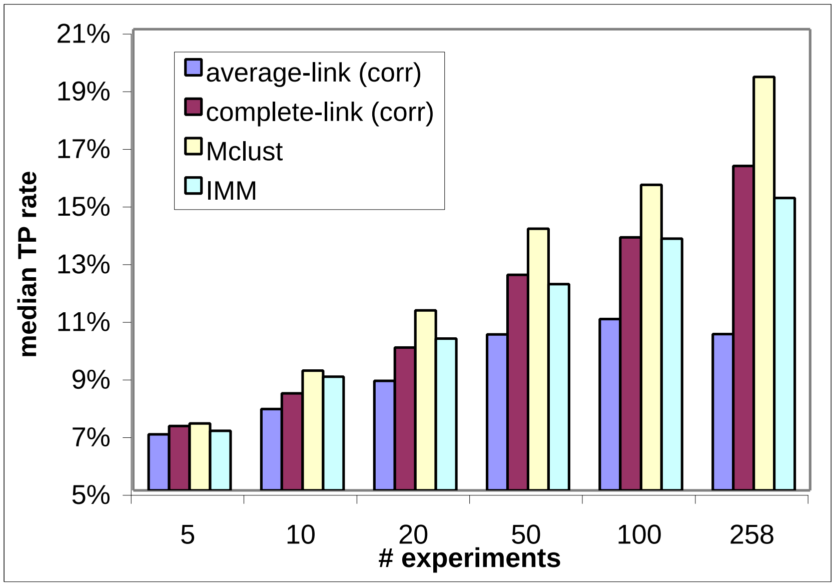


Figures E.1.b: Using ChIP data as the evaluation criterion, we compared the median TP rates using different numbers of experiments (E) from different clustering algorithms *at 25 clusters* on the compendium data subset with 215 genes. The median TP rates generally increase as E increases over different numbers of clusters.

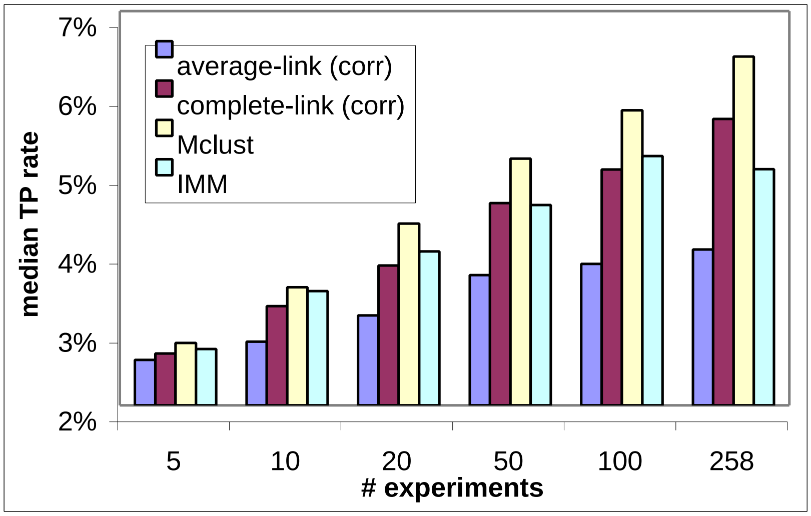


2. Compendium data subset with 537 genes and 258 experiments

Figures E.2.a: Using YPD as the evaluation criterion, we compared the median TP rates using different numbers of experiments (E) from different clustering algorithms *at 40 clusters* on the compendium data subset with 537 genes. The median TP rates generally increase as E increases over different numbers of clusters.



Figures E.2.b: Using ChIP data as the evaluation criterion, we compared the median TP rates using different numbers of experiments (E) from different clustering algorithms *at 40 clusters* on the compendium data subset with 537 genes. The median TP rates generally increase as E increases over different numbers of clusters.



TP and FN on ChIP vs YPD

In light of the comments from an anonymous reviewer, we explored the TP and FN rates from clustering results evaluated using YPD compared to the TP and FN rates evaluated using ChIP data.

Define:

$$TP\ rate = \frac{\text{\# gene pairs from the same clusters and share at least one common transcription factor}}{\text{\# gene pairs from the same clusters}}$$

$$FN\ rate = \frac{\text{\# gene pairs from different clusters but share at least one common transcription factor}}{\text{\# gene pairs sharing at least one common transcription factor}}$$

Let us consider applying the *complete-link* algorithm using correlation on the compendium data with 537 genes and 258 experiments:

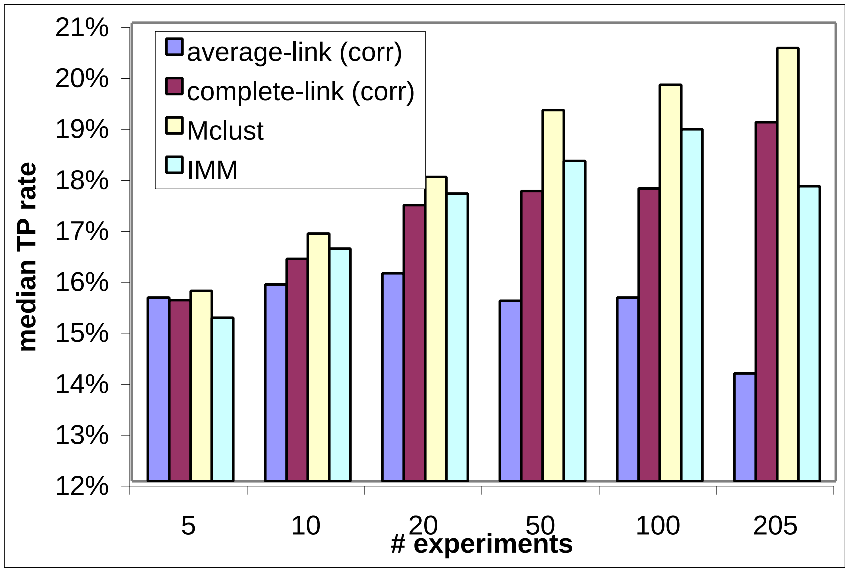
Evaluation criteria	TP rate	FN rate
YPD	16.3%	89.2%
ChIP	5.6%	90.3%

Note that the number of gene pairs sharing a common transcription factor on the ChIP data is significantly smaller than that using YPD. In particular, there are 7029 gene pairs sharing a common transcription factor using YPD. On the other hand, there are only 2710 gene pairs sharing a common transcription factor using a p-value of 0.001 on the ChIP data.

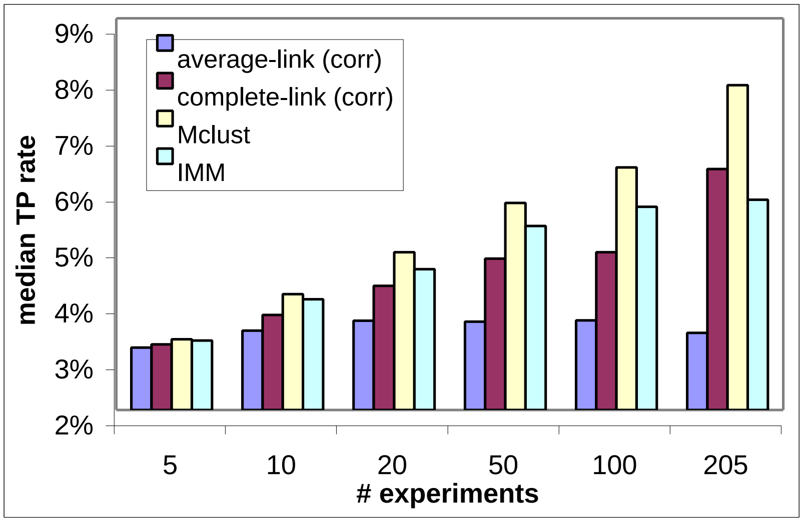
To summarize, we observed different TP and FN rates depending on the evaluation method.

3. Environmental stress data subset with 205 genes and 205 experiments

Figures E.3.a: Using SCPD as the evaluation criterion, we compared the median TP rates using different numbers of experiments (E) from different clustering algorithms *at 20 clusters* on the environmental stress data subset with 205 genes. The median TP rates generally increase as E increases over different numbers of clusters.

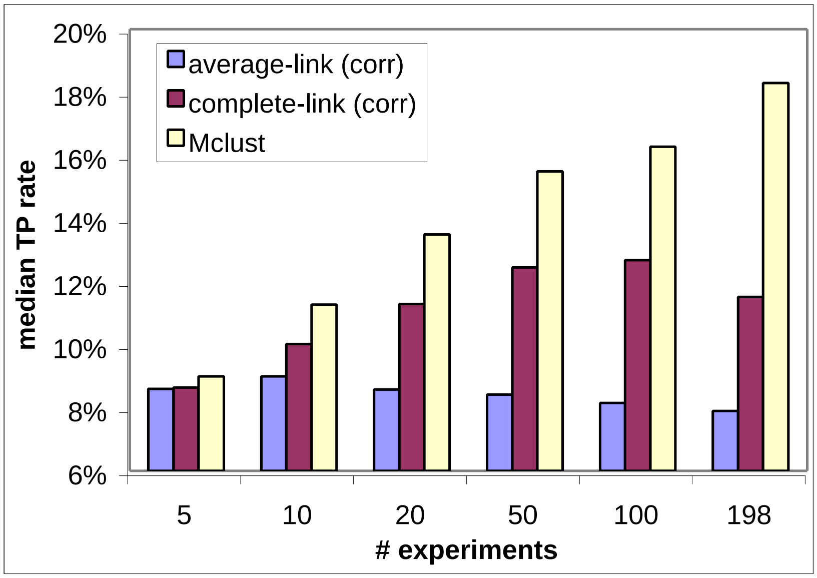


Figures E.3.b: Using ChIP data as the evaluation criterion, we compared the median TP rates using different numbers of experiments (E) from different clustering algorithms *at 20 clusters* on the environmental stress data subset with 205 genes. The median TP rates generally increase as E increases over different numbers of clusters.

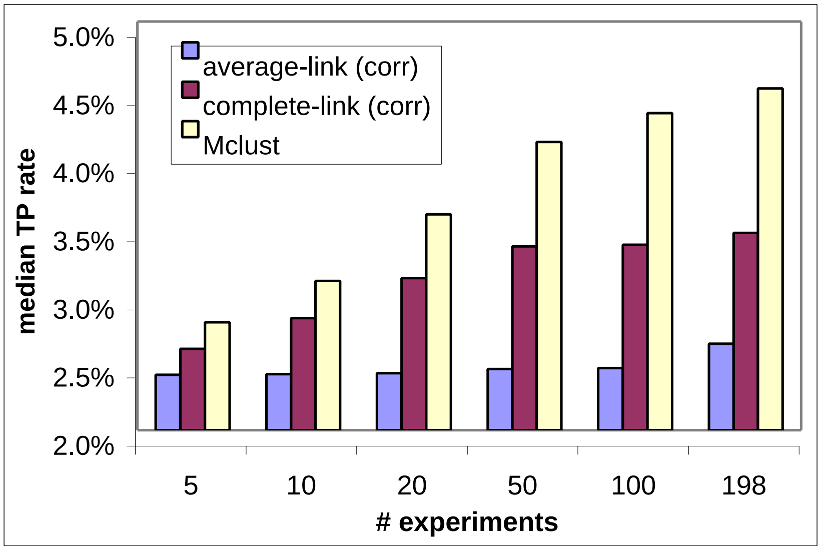


4. Environmental stress data subset with 526 genes and 198 experiments

Figures E.4.a: Using YPD as the evaluation criterion, we compared the median TP rates using different numbers of experiments (E) from different clustering algorithms *at 30 clusters* on the environmental stress data subset with 526 genes. The median TP rates generally increase as E increases over different numbers of clusters.



Figures E.4.b: Using ChIP data as the evaluation criterion, we compared the median TP rates using different numbers of experiments (E) from different clustering algorithms *at 30 clusters* on the environmental stress data subset with 526 genes. The median TP rates generally increase as E increases over different numbers of clusters.



F. Comparing ChIP data to YPD

Table F: We compared the bindings of gene transcription factor detected by ChIP data (Lee et al. 2002) to the known gene transcription factor interaction documented in YPD (Costanzo et al. 2000).

The total number of gene transcription factor interactions documented in YPD (as of Nov 2001) in the gene subset we considered is 791. The true positive percentage (TP %) is defined as the fraction of gene transcription factor interactions detected by both ChIP data and documented in YPD over all 791 interactions reported in YPD. The true negative percentage (TN %) is defined as the fraction of gene transcription factor interactions documented in YPD but not detected by ChIP data over all 791 interactions reported in YPD.

p-value	# TP (detected by both ChIP and YPD)	# FN (detected by YPD but not ChIP)	# detected by ChIP but not in YPD	TP %	FN %
0.001	159	632	421	20.1	79.9
0.005	196	595	775	24.78	75.22
0.010	218	573	1112	27.56	72.44
0.050	278	513	2759	35.15	64.85
0.100	318	473	4722	40.2	59.8

References

Costanzo, M.C., J.D. Hogan, M.E. Cusick, B.P. Davis, A.M. Fancher, P.E. Hodges, P. Kondu, C. Lengieza, J.E. Lew-Smith, C. Lingner, et al. 2000. The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res* **28**: 73-6.

Lee, T.I., N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, et al. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**: 799-804.

Medvedovic, M., K.Y. Yeung and R. Bumgarner. 2004. Bayesian mixture model based clustering of replicated microarray data. *To appear in Bioinformatics*.

Yeung, K.Y., C. Fraley, A. Murua, A.E. Raftery and W.L. Ruzzo. 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**: 977-987.

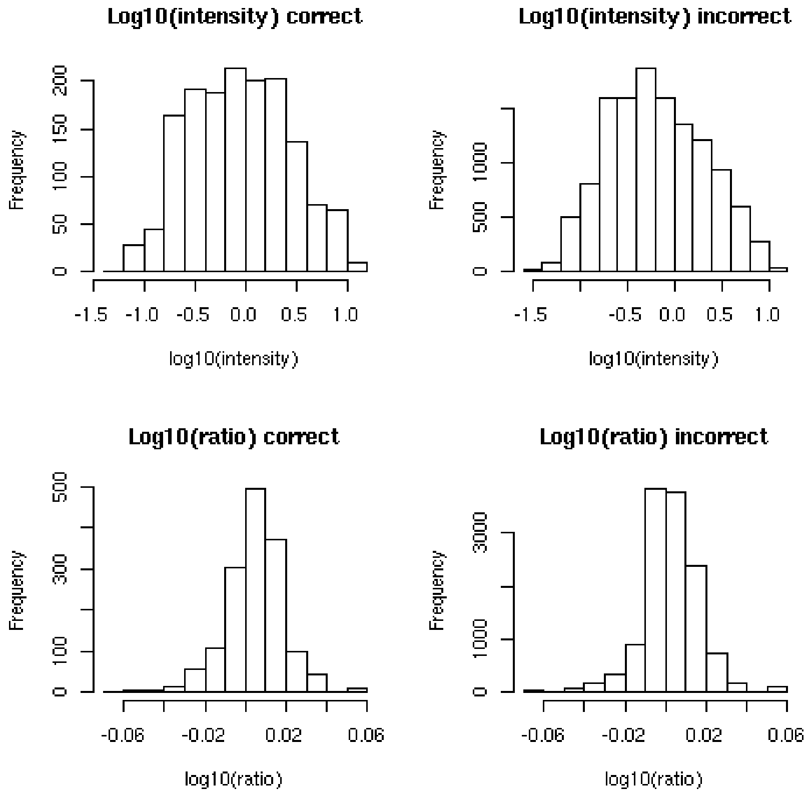
Yeung, K.Y., M. Medvedovic and R.E. Bumgarner. 2003. Clustering gene expression data with repeated measurements. *Genome Biology* **4**: R34.

G. Correct or incorrect classification of co-regulated genes

Effect of the thresholds of differential expression and absolute levels of expression
The gene subsets used in our experiments are selected based on the genes listed in the yeast transcription factor databases (SCPD and YPD). We did **not** pre-process the microarray data with any filtering steps based on differential expression or absolute levels of expression.

We explored the relationship between classification accuracy, expression ratio and expression intensity. Specifically, the log10(intensity) for each gene under each experiment is available from the yeast compendium data in addition to the log10(ratio). We are interested in the distributions of the log10(intensity) and log10(ratio) for correctly and incorrectly assigned gene pairs. In our exploratory study, correctly assigned gene pairs are assigned to the same cluster and share a common transcription factor according to YPD, while incorrectly assigned gene pairs share a common transcription factor but are assigned to different clusters. On the compendium data with 537 genes and evaluated using YPD, we plotted the following:

1) Histogram of the log10(**intensity**) for **correctly** assigned gene pairs
2) Histogram of the log10(**intensity**) for **incorrectly** assigned gene pairs
3) Histogram of the log10(**ratio**) for **correctly** assigned gene pairs
4) Histogram of the log10(**ratio**) for **incorrectly** assigned gene pairs



The above figure represents typical results over other data subsets. We observed that the shape of the histogram of the absolute level of expression (log10(intensity)) for the correctly assigned genes are not significantly different from that for the incorrectly assigned genes. We also observed similar results for the thresholds of expression ratio (log10(ratio)): the shape of the histogram of correctly assigned genes is not significantly different from that of incorrectly assigned genes pairs.

Comparing the distribution of mis-classified genes when the number of experiments is increased

Some genes that are correctly assigned when the number of experiment is small do get mis-classified when the number of experiments is increased. However, the proportion of such genes is smaller than those genes that become correctly assigned when the number of experiments is increased. This is generally expected since cluster groupings are not perfectly stable with changes in the number of experiments.